The Next Revolution in Computer Systems Architecture

Richard Oehler
Corporate Fellow
Office of the CTO

Computer Systems Architecture

AMD
Smarter Choice

Not just the Processor Chip

It's all the Chips and Interconnects

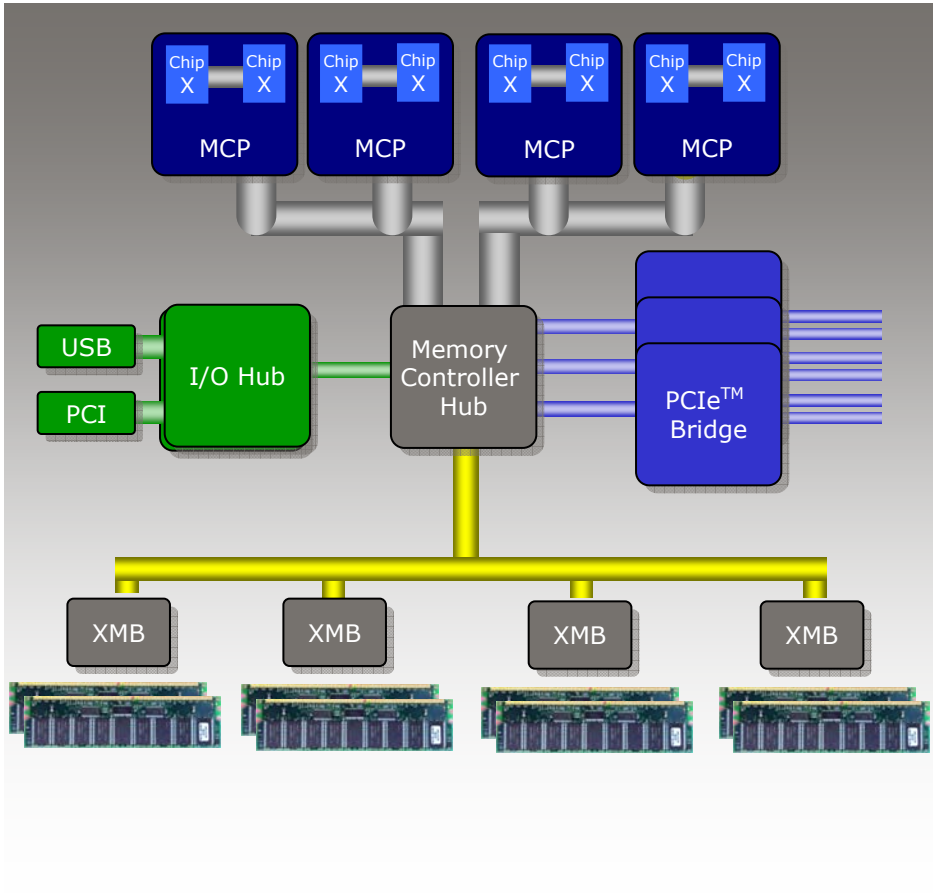
Chipsets, Memory, SMP Fabric, I/O, …

It's the Packaging

Form Factors, Power/Cooling

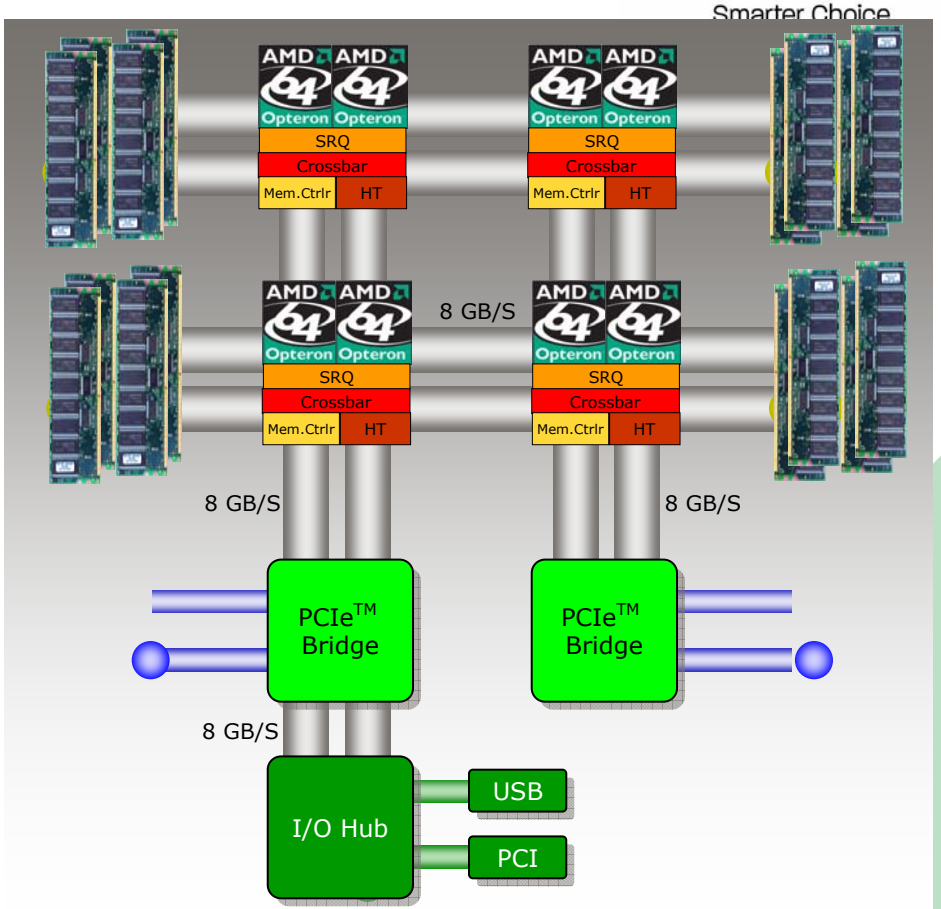
It's the Total System

What's Been Happening



Legacy x86 Architecture

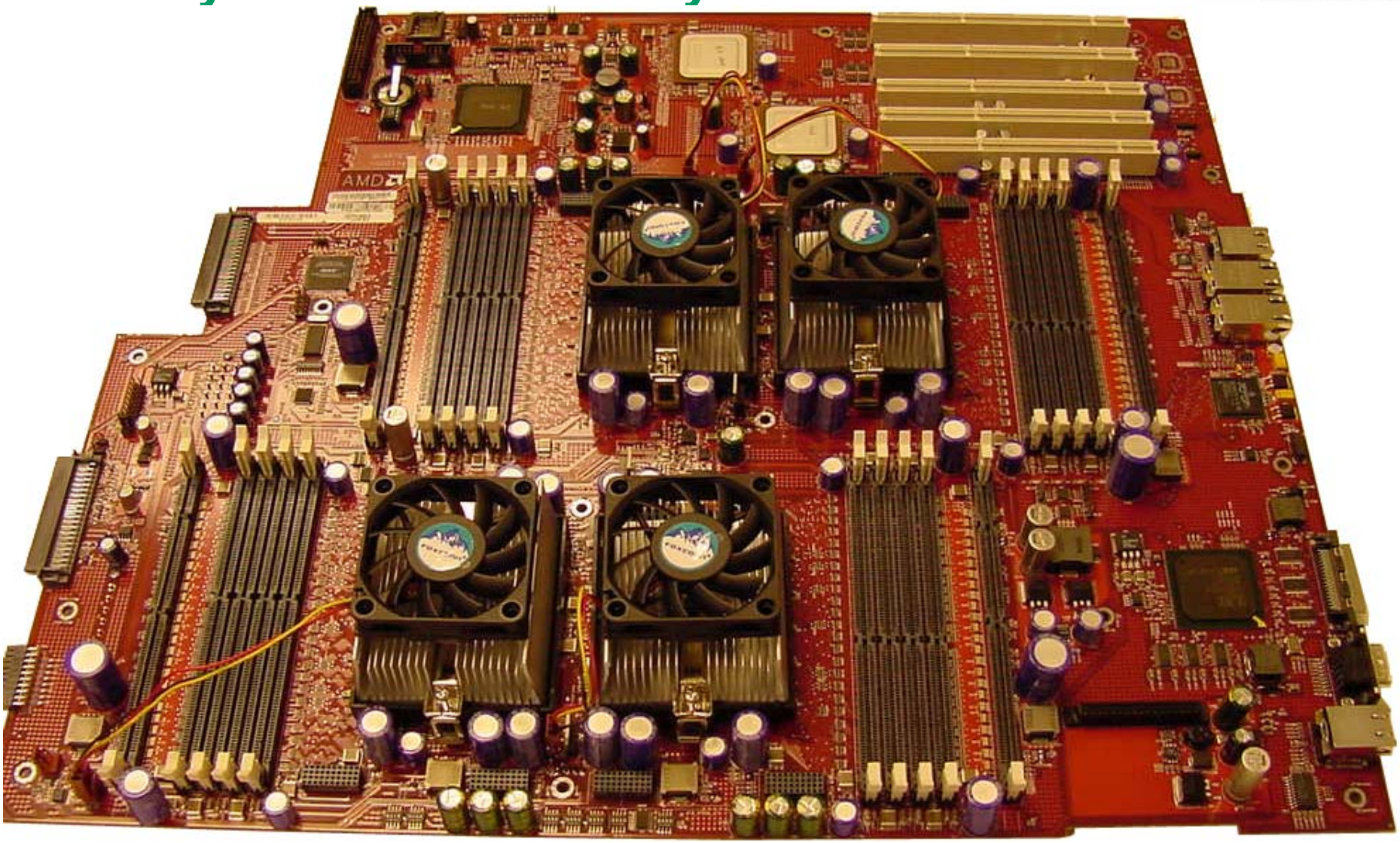
- 20-year old traditional front-side bus (FSB) architecture
- CPUs, Memory, I/O all share a bus
- Major bottleneck to performance
- Faster CPUs or more cores ≠ performance



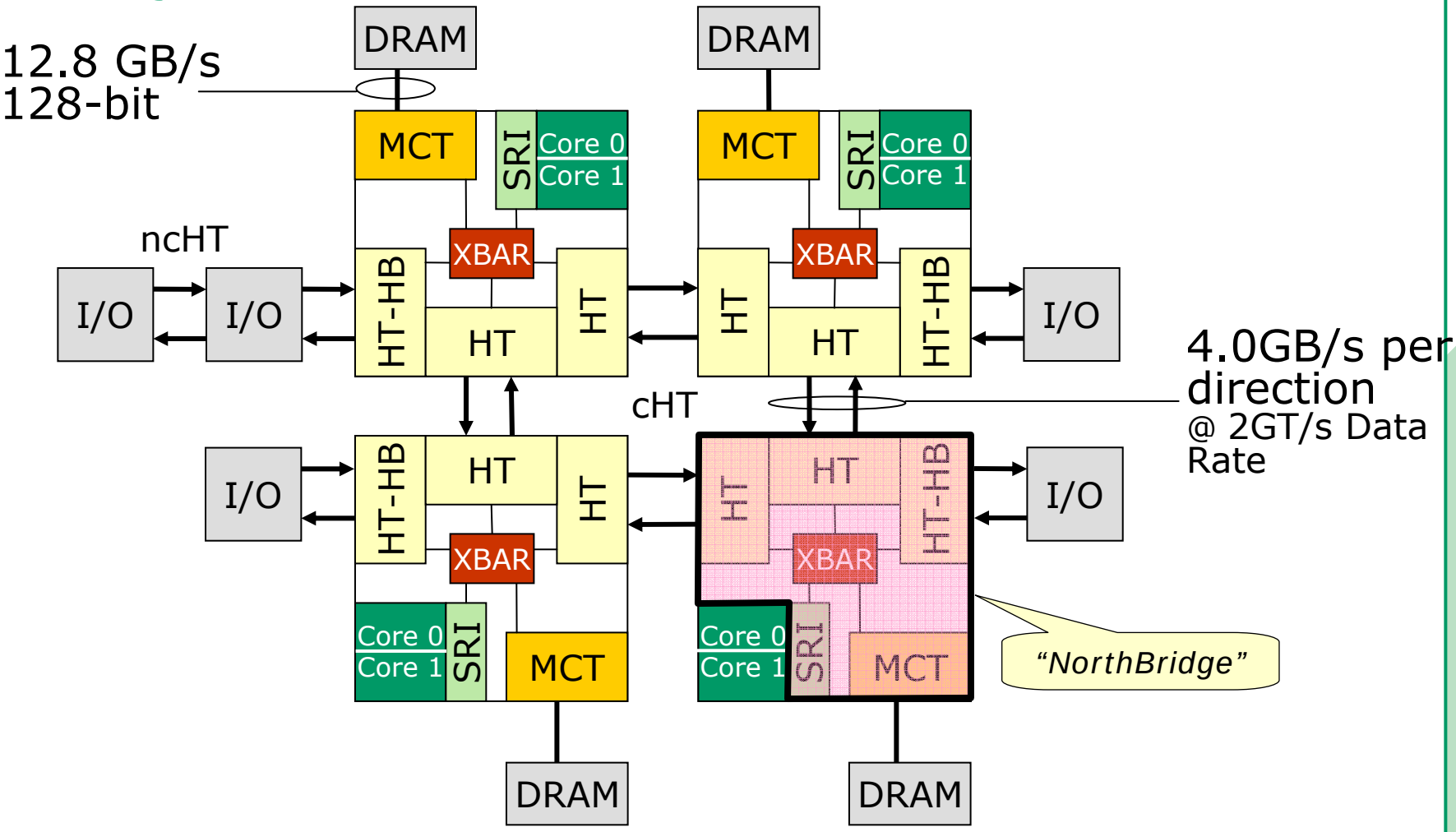
AMD64's Direct Connect Architecture

- Industry-standard technology
- Direct Connect Architecture reduces FSB bottlenecks
- HyperTransport™ interconnect offers scalable high bandwidth and low latency
- 4 memory controllers – increases memory capacity and bandwidth

4P System — Board Layout



AMD's Building Blocks Today, Tomorrow and the Future



Lessons Learned #1

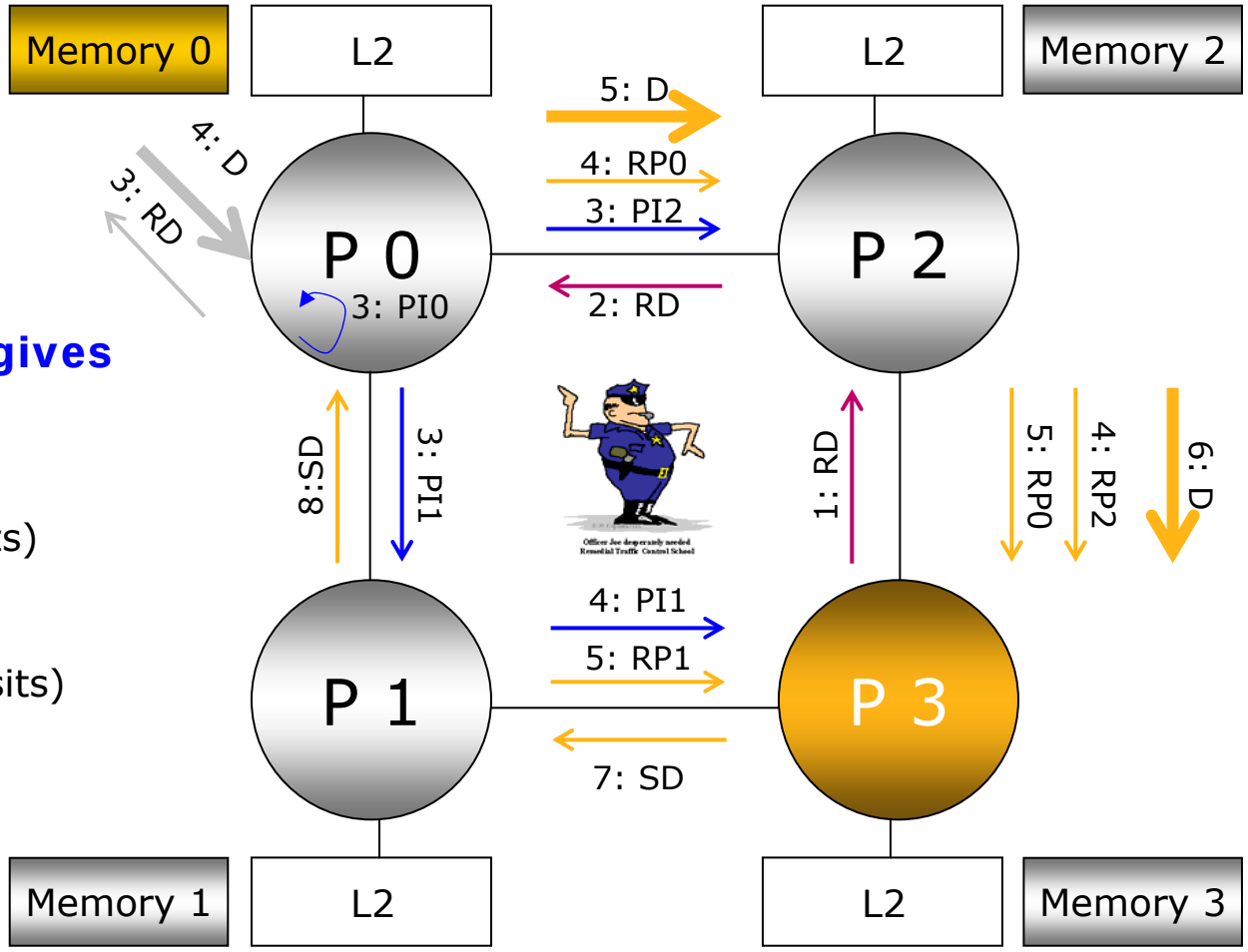
Allocation of XBAR Command buffer across Virtual Channels can have big impact on performance

AMD Smarter Choice

MP traffic analysis gives the best allocation

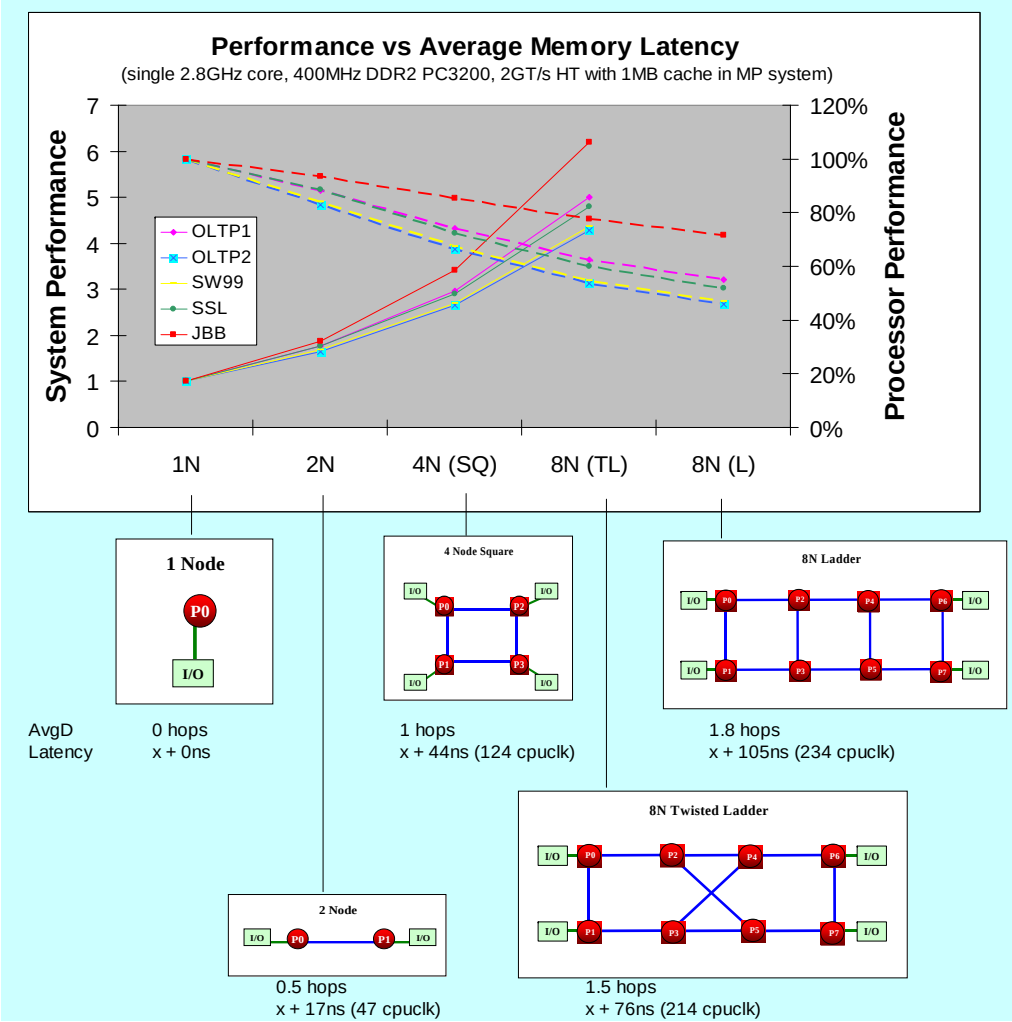
e.g. Opteron Read Transaction

- Request (2 visits)
- Probe (3 visits)
- Response (8 visits)



Lessons Learned #2

Memory Latency is the Key to Application Performance!



What's About To Happen

"Barcelona"…

Native quad-core upgrade for 2007

AMD Smarter Choice

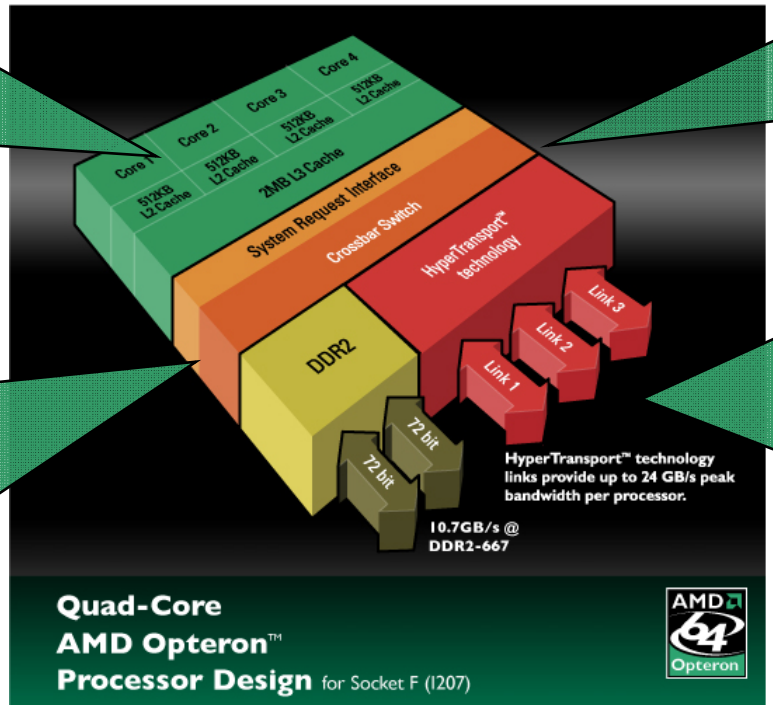
Native Quad-Core Processor

To increase performance-per-watt efficiencies **using the same Thermal Design Power.**

Advanced Process Technology

65nm Silicon-on Insulator Process

Fast transistors with low power leakage to **reduce power and heat.**



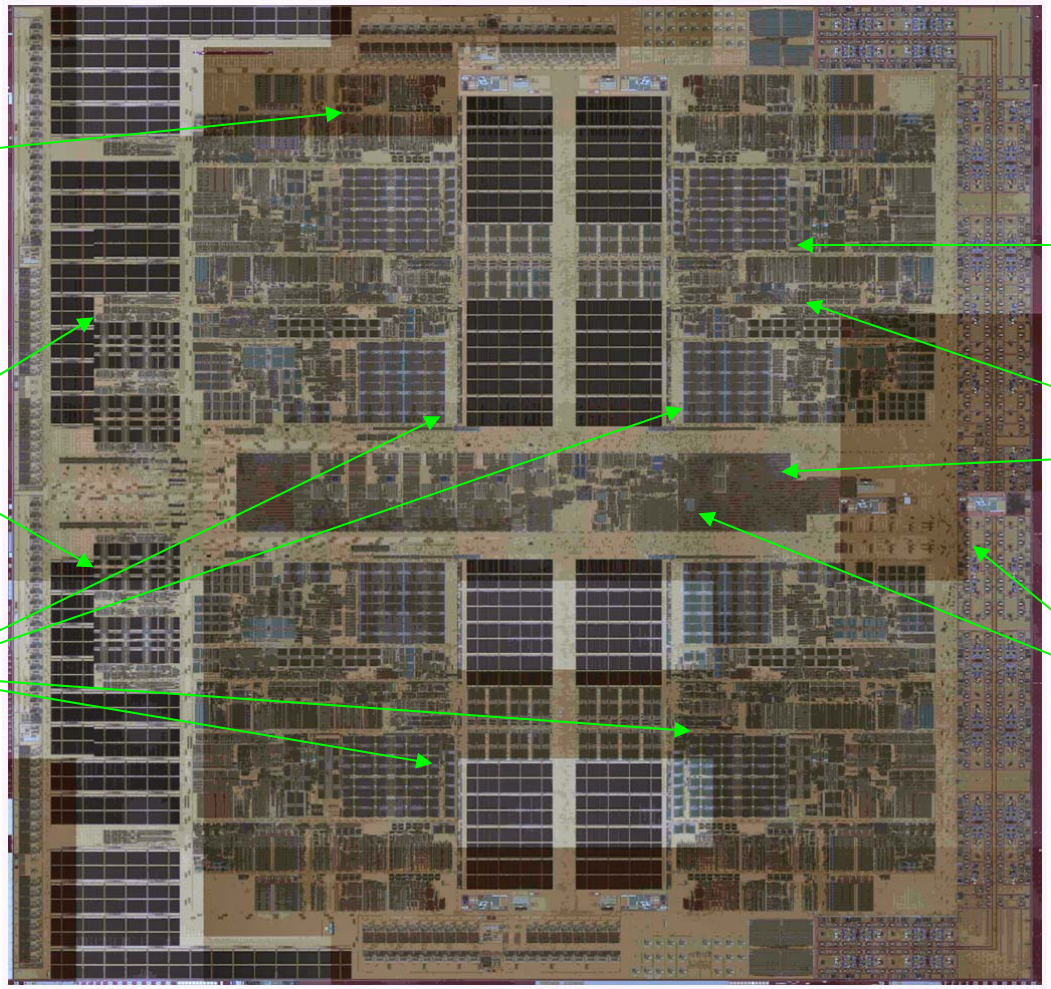
Platform Compatibility

Socket and thermal compatible with "Socket F".

Direct Connect Architecture

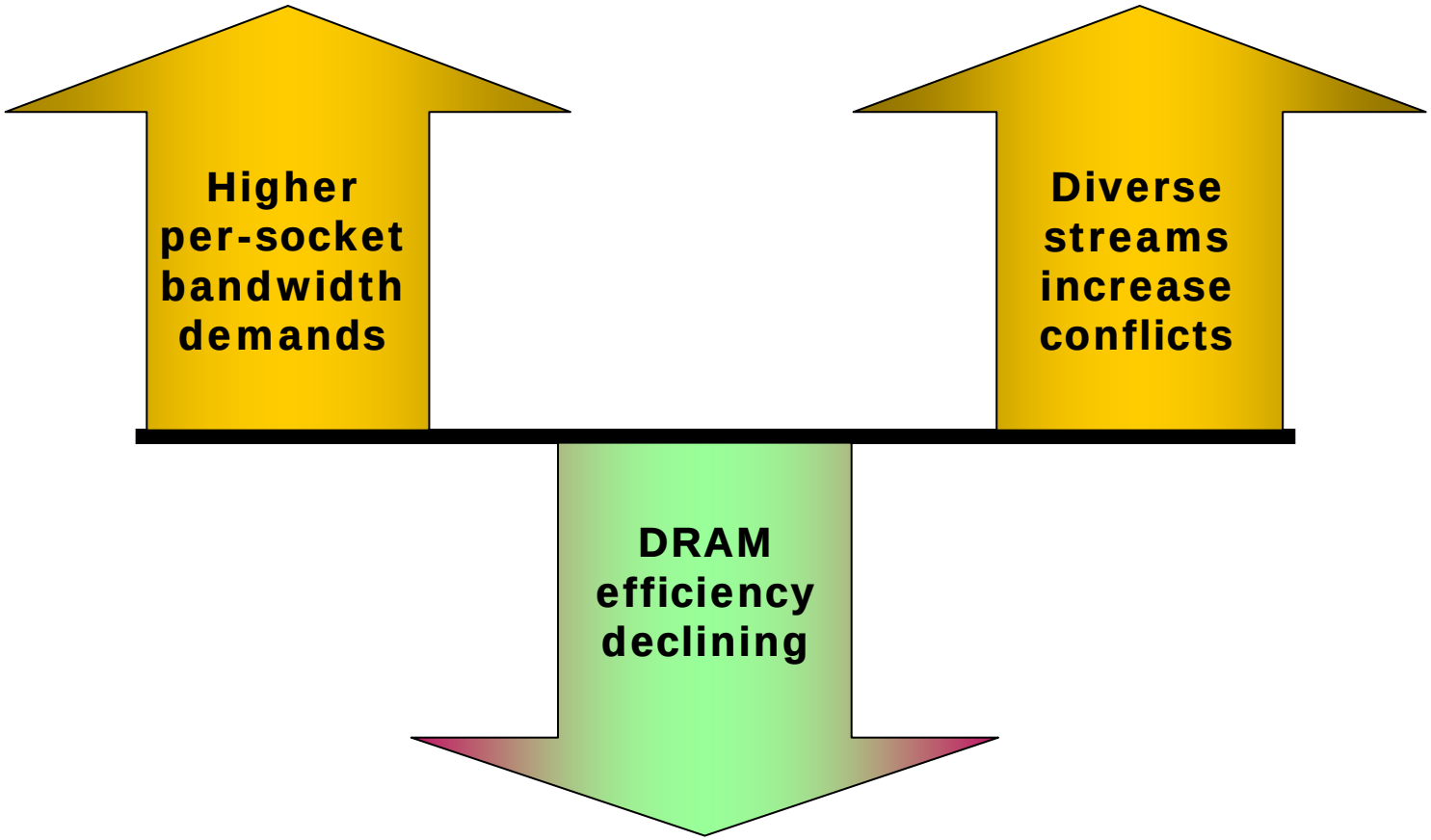
- Integrated memory controller designed for **reduced memory latency and increased performance**
  - Memory directly connected
- Provides **fast CPU-to-CPU communication**
  - CPUs directly connected
- **Glueless SMP** up to 8 sockets

The Barcelona Processor (4 core / die)



Trends in DRAM bandwidth

Improved Efficiency is the Answer



We must improve ***delivered*** DRAM bandwidth

Delivering more DRAM bandwidth

Independent DRAM controllers

Optimized DRAM paging

Re-architect NB for higher BW

Write bursting

DRAM prefetcher

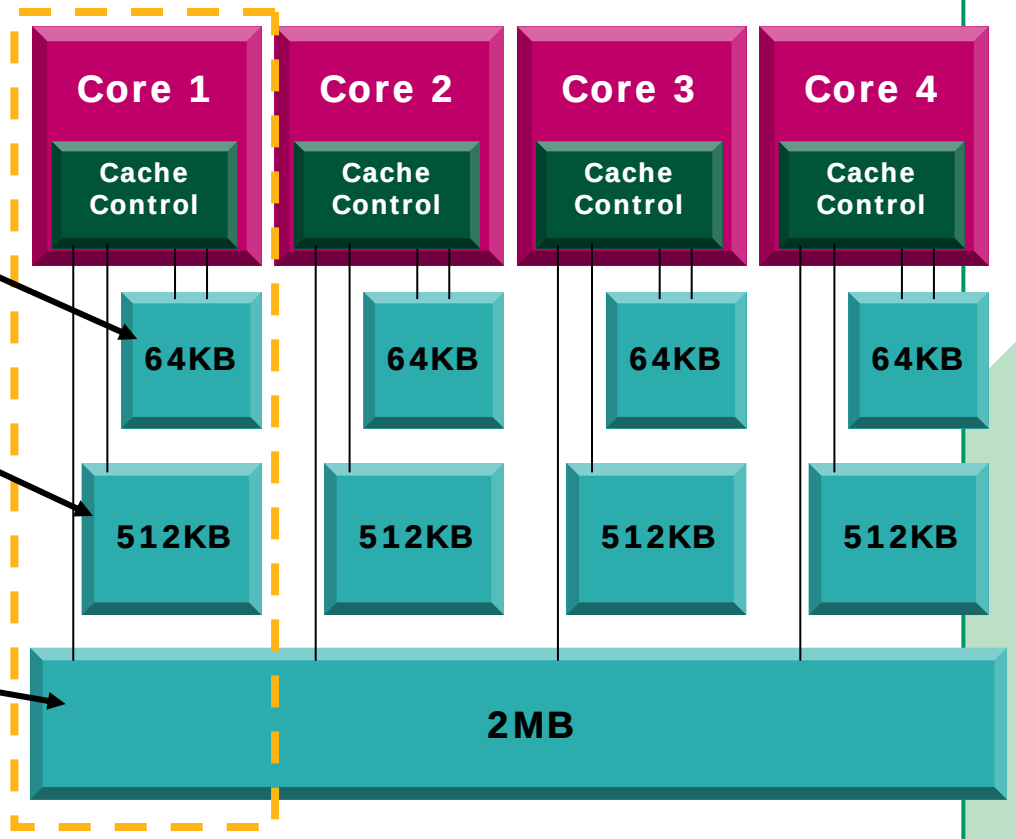
Core prefetchers

Balanced, Highly Efficient Cache Structure

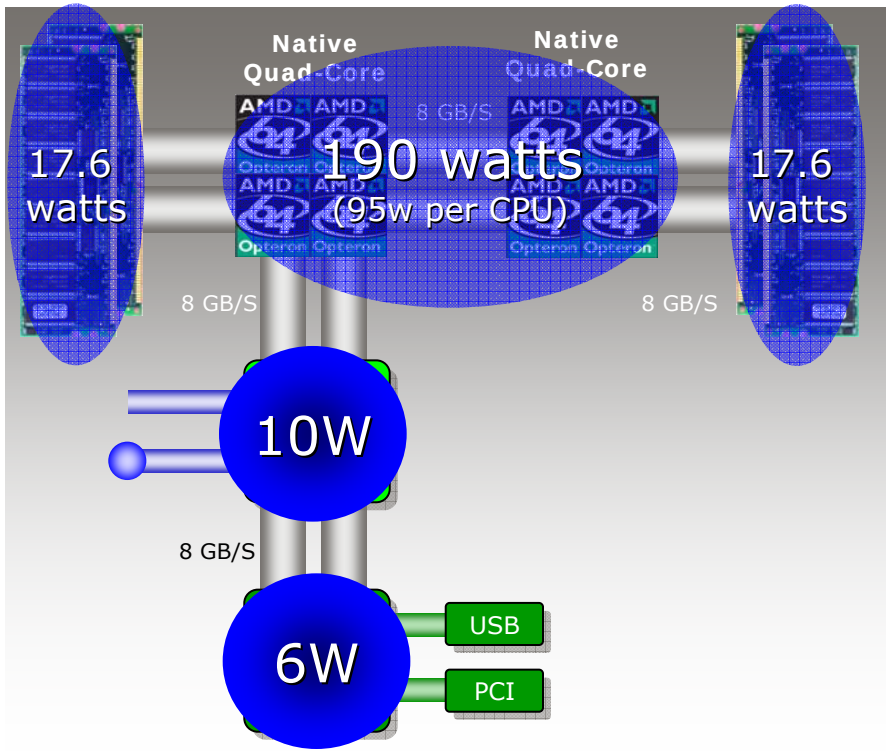
Dedicated L1
- Locality keeps most critical data in the L1 cache
- Lowest latency
- 2 loads per cycle

Dedicated L2
- Sized to accommodate the majority of working sets today
- Dedicated to eliminate conflicts common in shared caches
  - Better for Virtualization

Shared L3 – NEW
- Victim-cache architecture maximizes efficiency of cache hierarchy
- Fills from L3 leave likely shared lines in the L3
- Sharing-aware replacement policy
- Ready for expansion at the right time for customers



Quad-core System Power



2P System

- 190 watts for processors
- 16 watts for chipset
- 35.2 watts for DDR2

- Direct Connect Savings:
- *No external memory controller – saves 25 watts*
- *No FBDIMM – saves 48 watts*

System power is the metric that matters to our customers.

Direct Connect helps reduce system power.

Additional HyperTransport™ Ports

Enable Fully Connected 4 Node (four x16 HT) and 8 Node (eight x8 HT)

Reduced network diameter

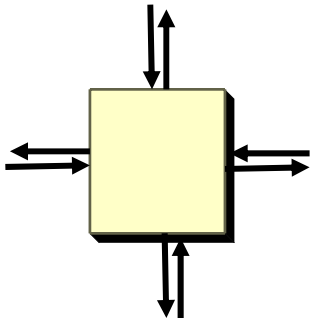
- Fewer hops to memory

Increased Coherent Bandwidth

- more links
- cHT packets visit fewer links
- HyperTransport3

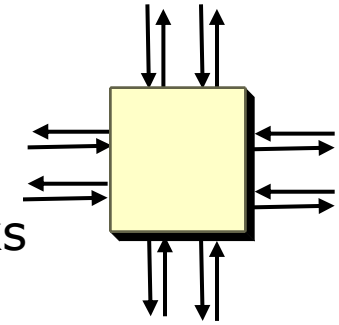
Benefits

- Low latency because of lower diameter
- Evenly balanced utilization of HyperTransport links
- Low queuing delays



Four x16 HT

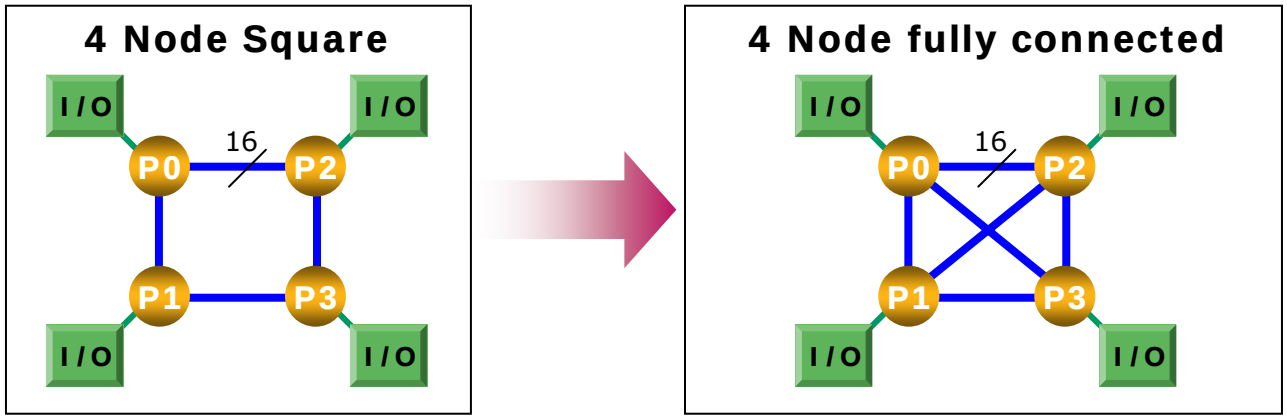
OR



Eight x8 HT

Low latency under load

4 Node Performance



4N SQ (2GT/s HyperTransport)

Diam 2 Avg Diam 1.00

XFIRE BW 14.9GB/s

+ 2 EXTRA LINKS

4N FC (2GT/s HyperTransport)

Diam 1 Avg Diam 0.75

XFIRE BW 29.9GB/s

(2X)

W/ HYPERTRANSPORT3

4N FC (4.4GT/s HyperTransport3)

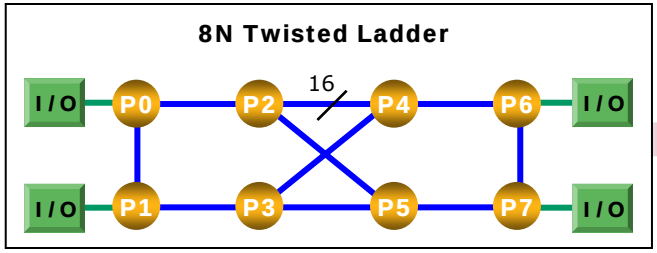
Diam 1 Avg Diam 0.75

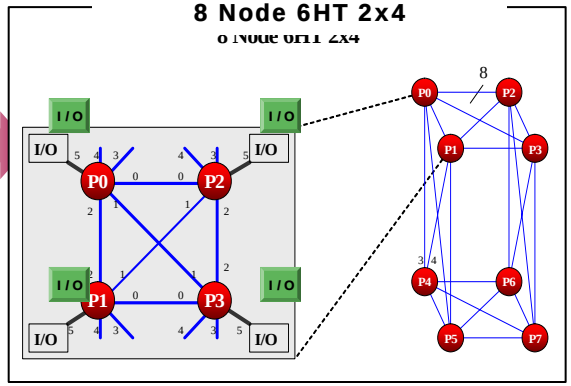
XFIRE BW 65.8GB/s

(4X)

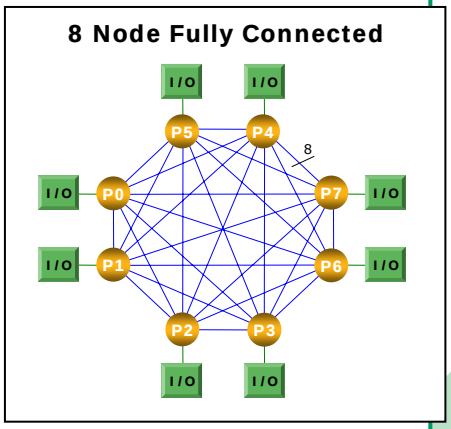
XFIRE ("crossfire") BW is the ***link-limited*** all-to-all communication bandwidth (data only)

8 Node Performance





OR



8N TL (2GT/s HyperTransport)

Diam 3 Avg Diam 1.62

XFIRE BW 15.2GB/s

8N 2x4 (4.4GT/s HyperTransport3)

Diam 2 Avg Diam 1.12

XFIRE BW 72.2GB/s

(5X)

8N FC (4.4GT/s HyperTransport3)

Diam 1 Avg Diam 0.88

XFIRE BW 94.4GB/s

(6X)

Interesting Questions

What About Hardware Multi-threading

AMD
Smarter Choice

Most implementations today

- Don't achieve consistent results
  - Some programs work well, some not so well but OK, some work well only with it turned off
  - Varies by with an application that contains many programs
  - Worse yet, varies within a program based on data input.
- Doesn't justify hardware costs
  - Unless the hardware is targeted at a specific market that is known to benefit from multi-threading
  - Increased complexity, chip area, schedule, risk

Given uncertain nature of benefit, becomes difficult for software to manage

- Always On or Off?
- Selectively On, based on program, or program data or …

But there is hope

- Carefully analysis, upfront SMT design, best practices engineering can lead to real benefits

Going Beyond Four cores

Opteron's fundamental design continues to scale

Relatively easy to get to beyond four cores

- Multi chip packaging v. single chip
- Next generation of process technology (45nm) will get us most way there

Will individual cores of the same design be faster in next generation process technology?

- Expect somewhat but not as much as previous generations
  - Circuit Stability v. Power/thermal limits

What about caches?

- Need for larger L2, L3
  - Not just Scaling effects, but memory bandwidth limitations
- Issues in SRAM design
  - Reliability - More transistors per cell (going from 6T to 8T or more)
- Alternative technology readiness
  - ZRAM, E-DRAM

And Power?

- Even at slower speeds, Processor scaling is still good

Looking Beyond 8 Cores

Challenges Going Beyond 8 Cores

AMD Smarter Choice

Sufficient Memory Bandwidth

Sufficient IO Bandwidth

Single thread performance

Extracting and managing multi-threading

Power/Cooling

Heterogeneous Cores

- Asymmetric Homogeneous Cores

Chip Level Packaging

Cost/Benefit

Yield & RAS

- Increase in Soft Errors as feature size continues to shrink
  - More exposure in Logic than Memory

Sufficient Memory Bandwidth

AMD has optimized memory bandwidth with many advanced techniques for current Opterons

- Both to reduce usage (e.g.. write combining) and to use empty cycles (e.g.. prefetching)
  - Law of diminishing returns

Require a balanced design

- ~ 30% loads/stores in typical instruction mixes
- Increasing number of cores increases total IPC by IPC of added cores*scaling
  - Current Designs are balanced for commercial and scientific workloads
- Reduced scaling (by a lot) if cores are not balanced with memory bandwidth

How to get more memory bandwidth

- More memory channels, Wider memory channels
  - Pin/Wiring issues for more/wider channels
- Increasing memory speed means reduced number of DIMMs
  - Or has significantly longer latency (FBDIMM)
- Stacked DIMMs may be a partial solution

More bandwidth with much longer latency is not the right answer

- OK for streaming date, not good for more random access

Many examples where significantly increasing cache improves cache hits thereby reducing overall memory bandwidth

- When is it better to use larger caches and less core
  - Application specific or not?

Sufficient I/O Bandwidth

AMD Smarter Choice

Opteron has really good I/O

- Best of breed I/O Bandwidth and Connectivity

Require a balanced design

- Rules of thumb
  - 1b i/o per instruction executed (commercial workloads - 70's,80's)
  - 1.2-1.4b i/o per instruction executed (more graphic content – 90's)
  - 1.6-1.8b i/o per instruction executed (managed code 00's)
- Assuming 1Giga Instructions/Sec, 8 cores will require between 1GB/sec and 1.8 GB/Sec sustained I/O transfers
- Now consider the efficiencies of realizing that bandwidth
  - I/O turn around time
  - Average size of packets
  - Overhead bytes v. payload
  - I/O channel loading
- 8 cores needs between 1.5 GB/sec and 5 GB/sec

Today's best I/O (PCI- Express realizes 2.5 to 5/GB/sec on 8/16 lanes

Adding SMT and pushing the number of cores x2 or x4 either

- Adds more i/o ports per chip or reduces inter-chip connectivity

Massive Connectivity (think TPC-C like environments) reduces efficiency of i/o

- Need more i/o ports based on nature of workload

Single Thread Performance

AMD
Smarter Choice

Single thread performance still matters

- Most current metrics are based on it
  - Limited multi-thread benchmarks

Many environments are not highly multi-threaded

- Digital Media
- Client Space
- Dusty Decks

Bulk of market does not have high degrees of multi threading

- How many design points are need to cover the major markets?
- In what order of introduction ?

Needed: more really parallel environments along with tools to mange same

Extracting and managing multi-threading

AMD
Smarter Choice

Some applications do not multi tread well

- Either not expressed well
- Or have limit inherent parallelism

Many applications have good multi thread potential

- Not currently organized for multi threading
  - Lack of/poor compilers, debuggers, OS support, tools for multi threading
- Writing correct and efficient multi thread applications is hard
  - Some estimates indicate less than 2% of the programming population can do it well

Very limited tools to find and extract automatically algorithm or non-coded parallelism

- One of the hardest problems in computer science
  - Been worked for the last 15-20 years with very little success

Power/Cooling

AMD Smarter Choice

Limits on power/cooling is how we got to multi-core to begin with

- Caused major rethink in how processors are designed
  - Contribute a significant percent to overall box power
  - Multi Chip makes problem worse
- Performance (or Price/Performance)/watt and Performance (or Price/Performance)/watt/cubic density are the new metrics

Pushing for a large number of cores/die while holding existing thermal envelopes results in slower cores

- New Designs save power in many ways
  - Selectively power reducing sections of cores or whole cores based on use
  - Overall performance limiting based on a maximum power consumption
- But core multiplier is still significant when each core consumes 3-5 watts

Not just a core/die problem, but a system problem

- Amount of total memory to provide a balanced system

Not just a system problem but a customer problem

- Physical limitations on providing more power/cooling
- Limited by various utility issues
- Power has become a serious TCO issue

Types of Cores: Homogeneous, Asymmetric, Sequester, Highly Reliable, …

AMD Smarter Choice

Today large number Homogeneous Cores/Threads are real issue for OS management

- Thread Dispatch queues are heavily contested
  - requiring new locking protocols and management
- Thread numbers bigger than OS design point
  - Number of threads packed in some word or double word structure
    - Major upheaval in OS
- Build it and they will come
  - Maybe not if overhead cancels expected improvement

Asymmetric Homogeneous Cores are an even more difficult problem

- Complicated hardware test and bring up
- Discovery and reporting to the OS
- Managing dispatching based on core type is non-trivial
  - Consider moving thread from slower to fast core based on new availability of faster core
  - Made even more complicated when balancing core power levels with overall application completion time

Types of Cores - continued

AMD Smarter Choice

Asymmetric Cores make matters worse

- Now must balance threads against different core types as well as all the previous issues

Sequested special purpose cores are a type of Asymmetric cores

- For your favorite special application or part thereof needs to run on a special
- Licensing Issues are non-Trivial
- Hidden from the OS
  - Accessed though device Drivers or Libraries or APIs
- Not just for software usage
  - Hardware prefetching, Hardware binary translation or Optimization
  - Opens up possible designs for high or very high reliability

Torrenza is an example of a asymmetric core

- Sequested or not
- Evolving from less efficient interconnect
- Expect to see Torrenza cores as an early first instance of asymmetric cores

Chip Level Packaging

Many degrees of freedom

- Larger chips v. Multi-Chip module
- Mixed cores and caches v. some cache on separate chip
- Pin count v. cost of package
- …

Can be used to put two (or more) muti-core chips together

- Often used to go the next level of multi cores
  - without having a full next level design
- Alternative structure separates some of cache hierarchy from cores
  - Significantly large, but slower access time caches

MCM works best if it uses an internal interconnect for local chips

- Not the standard external coherency interface
  - Internal NB interfaces
  - MCM Cores to MCM cores
  - MCM cores to L3 (or beyond) cache

Real cost breaks in increasing package/pin sizes

- Moving from plastic to ceramic is very non-linear
- Multi-core designs almost always need more pins, too many pins forces more specialized pin spacing and associated increasd in manufacturing and assembly cost

Different packages for different markets

- Acceptable cost v. market size
  - Volume is in digital media then client space

Cost / Benefit

AMD
Smarter Choice

As degree of multi core goes up

- What is level of scaling?
  - Can be very good if a balanced system is maintained
- If scaling falls off, design gets limited to more specialized applications
  - Economics of design, manufacturing change significantly as market contracts

What about modular designs?

- Highly structured
  - Mix and match
- Increased high level design time
- Most of design side cost is in debug/verification
  - Function of how many different actual designs
- Manufacturing cost increase
  - Different parts, SKUs
    - Demand prediction risks
    - Inventory management risks
  - Different manufacturing line optimizations
    - Especially with different die sizes
    - Low level tuning and process adaptation more complicated
      - Because there is more sizes

What are the cross over points?

- Market Economics,
- Available capital and resources

Yield and RAS

AMD Smarter Choice

As technology shrinks, density of circuits for a given die size significantly increases

- Defects more likely to make die less than perfect
  - Will need a partial good strategy
  - May need to use sparing at core level
    - Already done for some caches
- Increase in Soft Errors as feature size continues to shrink
  - More difficult in Logic than Memory
  - Error detection and correction and sparing common practice in caches
  - Until now Soft Logic Errors have not been a major issue
  - harder than memory errors to detect and correct
  - need new methodology and design tools

When is it better to design redundant cores or even TMR cores, than continue to add more and more detection and correction logic to individual cores?

- Yield curves, design complexity will determine cross over
- Redundant cores in specialized systems can be useful in markets that require very highly available systems

Should multi core chips be internally partitionable, especially in 24/7 environment?

- For service, diagnostics
- Where is cross-over between Scale out v. scale up?
  - Varies with application mix

How Big is Big Enough?

AMD
Smarter Choice

Why go bigger than 8?

- Consider 8 cores per die and 8 dies per system
  - 64 threads
- Add in SMT – number is at least doubled
- Position such a system against today's or near term largest SMP
  - Such multi-core/multi-chip systems are more powerful (by any measure) than what is currently in the market

Are there individual applications that need this much compute power?

- Biggest TPC-C benchmarks can be run easily on such an 8x8 a system
- Do real applications today or near future need this size?
  - Very few if any
    - As more cores are added, the number of applications requiring such power diminish non linearly

What about other environments?

- Server consolidation using virtualization scale can this large or larger if market demands it
  - But will need to carefully measure throughput v. reliability
- Hosted clients can be another such environment
  - Emerging market

What are the tradeoffs between increasing multi-core v. more chips per system?

- Economics

To go beyond 8 cores per die will require very reliable dies and systems

- MTBF terms are multiplied not subtracted

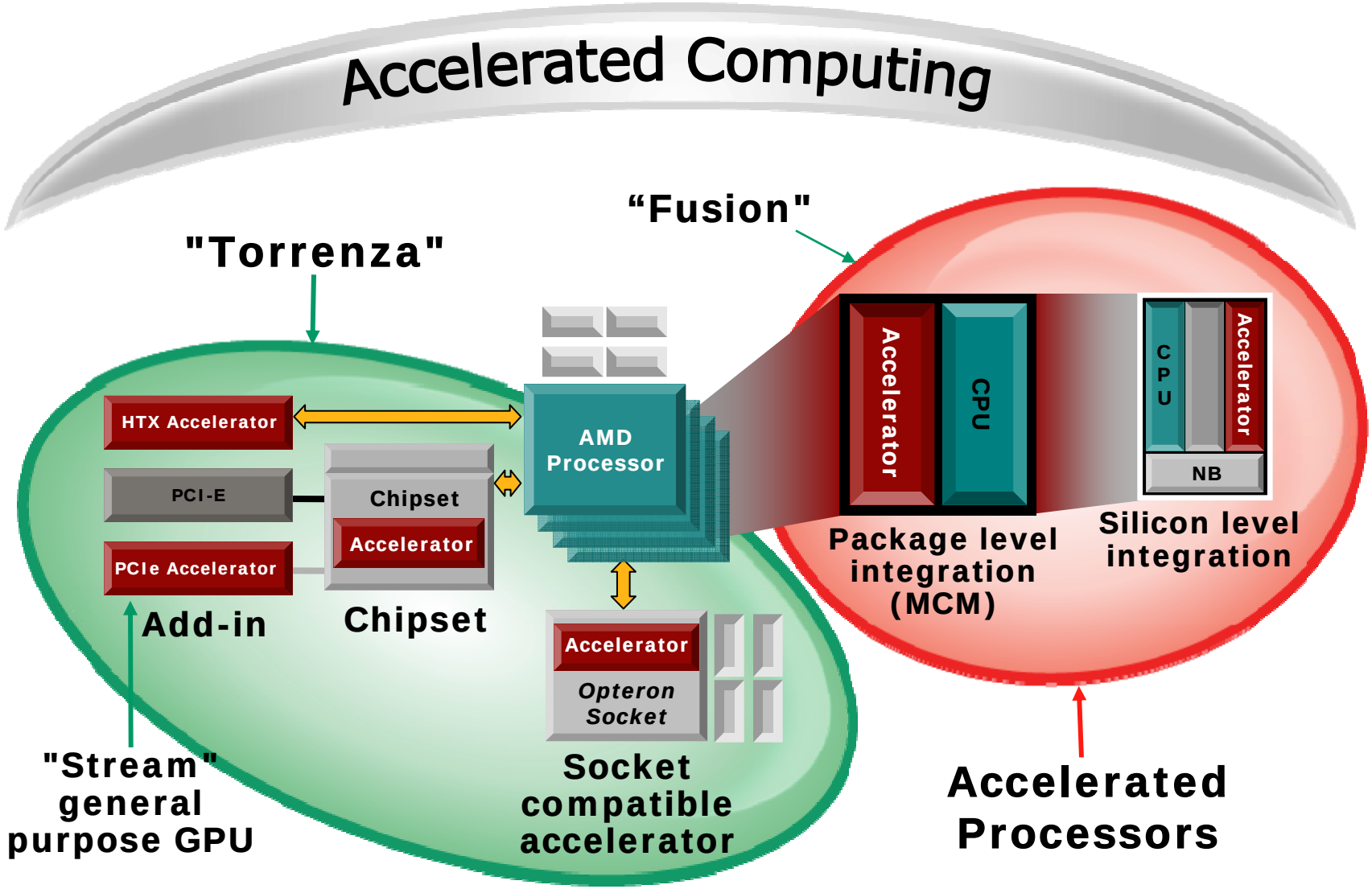
How does all of this relate to scale up?

- Good question
- Market leaning toward scale up
  - But with sufficiently large SMPs

AMD
Smarter Choice

The Next Big Things

Continuum of Solutions



AMD Smarter Choice

Torrenza & Fusion

First AMD Fusion Product: Accelerated Processor Combining CPU and GPU

AMD Smarter Choice

FUSION VISION

Create the optimal computing experience for an increasingly mobile, graphics- and media-centric world

Deliver improvements in microprocessor performance-per-watt-per-dollar over today's CPU-only architectures

Continue to scale x86 by enabling new x86 computing paradigms, classes and form factors

The Next Big Things

AMD
Smarter Choice

AMD, the company that first brought the x86 industry ...

Simultaneous 32-bit / 64-bit computing

The integrated memory controller

HyperTransport™ Technology

Native multi-core

The truly open x86 platform

Is once again changing the world of computing:

Torrenza and Fusion

Backup

AMD
Smarter Choice

Abstract

Computer Systems Architecture is going through a major re-thinking. Constraints from form factors, to scaling, to power/cooling, to system balance, have overwhelmed current designs. This talk will discus these reasons and a few others that have caused this to happen, what some of the new design ideas/parameters are, how they will manifest themselves in systems in the not too distant future and what more needs to be done before there is a new stable base.

Talk Outline

AMD
Smarter Choice

Computer Systems Architecture

Today's World

- Form Factors,
- Scaling,
- Power/Cooling,
- System balance,

Not So New Ideas

- Multi-core
- Accelerators
- Heterogeneous

Issues to be Solved

Future Directions

Delivering more DRAM bandwidth

AMD Smarter Choice

Independent DRAM controllers

Optimized DRAM paging

Re-architect NB for higher BW

Write bursting

DRAM prefetcher

Core prefetchers

Concurrency

More DRAM banks

- **reduces page conflicts**

Longer burst length

- **improves command efficiency**

Delivering more DRAM bandwidth

Independent DRAM controllers

Optimized DRAM paging

Re-architect NB for higher BW

Write bursting

DRAM prefetcher

Core prefetchers

Increase page hits, decrease page conflicts

History-based pattern predictor

Delivering more DRAM bandwidth

Independent DRAM controllers

Optimized DRAM paging

Re-architect NB for higher bw

Write bursting

DRAM prefetcher

Core prefetchers

Increase buffer sizes

Optimize schedulers

Ready to support future DRAM technologies

Delivering more DRAM bandwidth

Independent DRAM controllers

Optimized DRAM paging

Re-architect NB for higher BW

Write bursting

Minimize Rd/Wr Turnaround

DRAM prefetcher

Core prefetchers

Delivering more DRAM bandwidth

Independent DRAM controllers

Optimized DRAM paging

Re-architect NB for higher BW

Write bursting

DRAM prefetcher

- **Track positive and negative, unit and non-unit strides**
- **Dedicated buffer for prefetched data**
- **Aggressively fill idle DRAM cycles**

Core prefetchers

Delivering more DRAM bandwidth

Independent DRAM controllers

Optimized DRAM paging

Re-architect NB for higher BW

Write bursting

DRAM prefetcher

Core prefetchers

DC Prefetcher fills directly to L1 Cache

IC Prefetcher more flexible

- 2 outstanding requests to any address